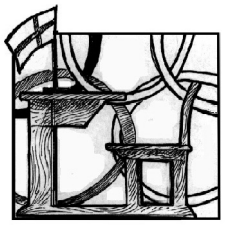

Perinnöllisyyttä ja tilastotiedettä

Mikko J. Sillanpää
Matemaattisten tieteiden laitos, Biologian laitos ja Biocenter Oulu
Oulun yliopisto
mikko.sillanpaa@oulu.fi

Modernin molekyylibiologian laboratoriotekniikat tuottavat suuria määriä geneettisiä mittausaineistoja. Tilastomatematiikan lineaarisia regressiomalleja voidaan käyttää kytkemään tutkittavan perinnöllisen ominaisuuden (kuten esimerkiksi ihmisen pituuteen, verenpaineeseen tai johonkin sairauteen liittyvät mittaukset) ihmisen DNA:sta tehtyihin mittauksiin. Tällöin voidaan saada tietoa, missä kohdassa genomia (l. kromosomistoa) olevat mittaukset sopivat parhaiten yhteen tutkittavan perinnöllisen ominaisuuden mittauksien kanssa. Tällaista toimenpidettä eli näkyvän tason havaintojen ”sovittamista” geneettisen tason mittauksiin (l. genotyyppeihin) kutsutaan yleisesti geenien kartoitustehtäväksi (ks. Esimerkki). Vaihtoehtoisesti samanlaista tilastomatematiikan mallia voidaan käyttää vaihtelevalla menestyksellä ennustamaan tutkittavaa perinnöllisen ominaisuuden arvoa tai sairastumisalttiutta/riskiä henkilöillä DNA:sta tehtyjen mittausten perusteella. Vastaavien tilastollisten mallien tutkimus ja käyttö modernissa kasvi- ja eläinjalostuksessa ennustamaan kyseisen lajikkeen tai eläimen perinnölliseen ominaisuuteen liittyvää jalostuksellista arvoa on yksi tämän hetken ajankohtaisimpia maatalouteen liittyviä tutkimuskysymyksiä (ks. Juga et al., 2012).

Koska useat perinnölliset ominaisuudet näyttäisivät olevan tämänhetkisen tutkimuksen valossa sellaisia, että niihin samanaikaisesti vaikuttaa suuri joukko geenejä ja ympäristöllisiä tekijöitä, on niiden paikantaminen ja käyttö ennustetarkoituksiin tehokkaampaa malleilla, jotka tarkastelevat useampaa mittauskohtaa samanaikaisesti. Toisaalta koska yksilöiden määrä, joilta DNA-mittaukset otetaan, on tyypillisesti paljon pienempi kuin mittauspisteiden määrä, ei jokaisen mittauspisteen vaikutusta voida samanaikaisesti arvioida (koska mahdollisten ratkaisujen määrä on suuri) ilman, että käytetään niin sanottua tilastollista muuttujanvalintaa tai *a priori* -informaatiota (esimerkiksi pakottamalla suuri joukko mittauspisteiden vaikutuksista nollaan).

Tällainen suurien mittausaineistojen tyypillinen ongelma tunnetaan nimellä ”small n, large p”, ja sen arviointiin käytettävät muuttujanvalinnan tilastolliset menetelmät ovat tällä hetkellä monessa perinnöllisyyteen liittyvässä tutkimuskysymyksessä keskeisessä asemassa. Tutkimusaineistoissa voi tyypillisesti olla satoja tuhansia tai jopa miljoonia mittauspisteitä pitkin DNA:ta, jotka on mitattu tyypillisesti sadoilta tai tuhansilta tutkimusyksilöiltä.

Tällaisten ongelmien tilastollisten ratkaisumenetelmien suunnittelu ja jatkokehitys on geneettisten aineistojen käsittelyyn erikoistuneiden tilastotieteen tutkijoiden arkipäivää. Tällaisia henkilöitä koulutettaessa on tilastotieteen, matematiikan ja sovelletun matematiikan opintojen luoma luja pohja sellainen kivijalka, jota on mahdoton muilla opinnoilla korvata. Toisaalta

myös perinnöllisyystiedettä pitää jaksaa opiskella niin paljon, että sen käsitteillä voi vaivatta operoida. Siksi aito monitieteisyys ei mainosarvostaan huolimatta ole laji, jossa saadaan nopeita voittoja ja huikeita valmistumisaikoja.

Henkilöitä, jotka käyttävät, ymmärtävät tai kehittävät yllä kuvatun kaltaisia tilastomatematiikan menetelmiä, on työmarkkinoilla jatkuvasti liian vähän kysyntään nähden. Tässä joukossa erityisesti ”hyvin biologiaa puhuvia” matematiikasta tai tilastotieteestä valmistuneita maistereita/tohtoreita on niukasti. Siksi haluankin suositella matematiikan ja tilastotieteen maisteriopintoja pohjaopinnoiksi biologiasta kiinnostuneille opiskelijoille.

Esimerkki tyypillisestä lineaarisesta regressiomallista on

$$y_i = b_0 + \sum_{j=1}^{p} x_{ij} b_j + e_i \quad (i = 1, \ldots, n).$$

Tässä y_i on näkyvän tason havainto tutkittavasta perinnöllisestä ominaisuudesta yksilöllä i, b_0 on vakiotermi, joka halutaan selvittää, x_{ij} on geneettisen tason mittausarvo (esimerkiksi -1 genotyypille AA, 0 genotyypille AB ja 1 genotyypille BB) yksilölle i DNA:n kohdassa j. Kulmakerroin b_j kuvaa mittauspisteen vaikutusta tutkittavaan ominaisuuteen kohdassa j, joka halutaan selvittää kaikissa DNA:n kohdissa. Jäännöstermit e_i oletetaan samoin jakautuneiksi ja keskenään riippumattomiksi siten, että ne kukin noudattavat samanlaista normaalijakaumaa, $e_i \sim \mathrm{N}(0, \sigma^2)$, varianssilla σ^2. Tämä jäännöstermeille tehty oletus antaa yleisesti kriteerin mallin sopivuuden tarkasteluun aineistossa $((y_i, x_{ij}), i = 1, \ldots, n; j = 1, \ldots, p)$, jonka perusteella myös tuntemattomille muuttujille voidaan tuottaa arviot. Jotta arviot voidaan tuottaa myös tilanteessa, kun $p > n$, lisäksi tarvitaan muuttujan valintaa tai *a priori* -informaatiota.

Muuta aiheesta suomenkielellä:

Juga, J., Sillanpää M. J., Mäntysaari E. (2012) ”Lypsykarjan genominen valinta” kirjassa: ”Maailma muuttuu: muuttuuko maatalous.” Sivut 165–172. Mervi Seppänen (ed.).

Verkko-Solmun oppimateriaalit

Osoitteesta `http://solmu.math.helsinki.fi/oppimateriaalit.html` löytyvät oppimateriaalit:

- Reaalianalyysiä englanniksi (William Trench)
- Geometrian perusteita (Matti Lehtinen)
- Geometria (K. Väisälä)
- Lukualueiden laajentamisesta (Tuomas Korppi)
- Jaksolliset desimaaliesitykset algebrallisesta näkökulmasta (Jaska Poranen ja Pentti Haukkanen)
- Algebra (Tauno Metsänkylä ja Marjatta Näätänen)
- Algebra (K. Väisälä)
- Matemaattista fysiikkaa lukiolaiselle (Markku Halmetoja ja Jorma Merikoski)
- Lukuteorian helmiä lukiolaisille (Jukka Pihko)
- Matematiikan peruskäsitteiden historia (Erkki Luoma-aho)
- Matematiikan historia (Matti Lehtinen)